Problem state：

1. 介绍近似方法（即将分析分布变为分析定值），并证明差值
2. 介绍该问题的greedy算法先计算greedy算法在不同的duration下的期望变化

Solution frame for bandit with static-batch arms：

1. 不同的固定batch在特定环境下（即确定duration）对应的reward，证明在特定的duration之下存在最优的固定的batch。1.先证特定batch对应的reward；2由上面的推论出每个特定的duration存在最优的batch。
2. 在1的情况下说明当duration变化时该问题转化为bandit问题。介绍bandit和折扣UCB
3. 性能分析：1. 计算opt的上界和下界。2.计算greedy算法的竞争比。3计算最优batch的竞争比。4计算discount-UCB的regret和竞争比。

Solution frame for bandit with DQN-arms：

1. 介绍在确定duration的基于batch的匹配过程可以视为马尔科夫过程
2. 证明在确定duration基于马尔科夫过程的最优算法应当优于最优的固定batch的方案
3. 介绍基于马尔科夫过程的强化学习和bandit结合的具体思路
4. 基于上面进行性能分析，竞争比
5. 介绍DQN